# Results on Local Stability of Fixed Step Size Recursive Algorithms

J.A. Bucklew*, T.G. Kurtz**, W.A. Sethares*

Department of Electrical and Computer Engineering*
Department of Mathematics**
University of Wisconsin-Madison, Madison, WI 53706 USA

## Abstract

A recursive equation which subsumes several common adaptive filtering algorithms is analyzed for general stochastic inputs and disturbances by relating the motion of the parameter estimate errors to the behavior of an unforced deterministic ordinary differential equation (ODE). Local stability of the ODE implies weak convergence of the algorithm while instability of the differential equation implies nonconvergence of the parameter estimates. The analysis does not require continuity of the update equation, and the asymptotic distribution of the parameter trajectories for all stable cases (under some mild conditions) is shown to be an Ornstein - Uhlenbeck process.

The ODE's describing the motion of several common adaptive filters are examined in some simple settings, including the Least Mean Square (LMS) algorithm and all three of its signed variants (the signed regressor, the signed error, and the sign-sign algorithms). Stability and instability results are presented in terms of the eigenvalues of a correlation-like matrix. This generalizes known results for LMS, signed regressor and signed error LMS, and gives new stability criteria for the sign-sign algorithm.

## 1. Introduction

As applications of adaptive filtering, communication, control, and identification methods have grown so have the number of adaptive algorithms. Some are proposed because of their convergence properties, some because of their numerical simplicity, and others because of their noise rejection capabilities. The general recursive form

$$W_{k+1} = W_k + \mu H(W_k, Y_k, U_{k+1}) \qquad (1)$$

captures most of these algorithms by suitable choice of $H(\cdot)$. In (1), $W_k$ represents the parameter estimate errors, $Y_k$ is (usually) a vector of inputs, $U_k$ is a disturbance process that represents all nonidealities such as measurement and modeling errors, and $\mu$ is a small positive constant stepsize. Convergence of the process $W_k$ to a stationary distribution about zero is equivalent to convergence of the adaptive filter parameter estimates to a region about their optimal values. Two important questions concerning the behavior of $W_k$ arise immediately: Under what conditions is the process stable? When do there exist stationary distributions for $W_k$, and how can these stationary distributions be characterized?

Let us define a time scaled continuous time version of (1) as

$$W_\mu(t) = W_{[t/\mu]} \qquad (2)$$

where $[z]$ represents the integer part of $z$.

We address our questions by relating the behavior of the scaled adaptive algorithm (2) for small $\mu$ to the behavior of the associated deterministic ordinary differential equation (ODE)

$$W(t) = W_0 + \int_0^t \hat{H}(W(s))ds \qquad (3)$$

where $\hat{H}(\cdot)$ is a smoothed version of $H(\cdot,\cdot,\cdot)$.

The question of when time scaled versions of the $W_k$ process converge (as $\mu \to 0$) to $W(t)$ has been investigated by a number of researchers both for fixed $\mu$ and for the time varying stepsize cases of stochastic approximation. Many of the original notions (done in the stochastic approximations context) are due to Ljung [3] though the present approach is probably closest in spirit to [2]. Excellent overviews of this area are available in the books [1], [2]. Our arguments are somewhat shorter than some of the more traditional pathways and we are able to prove the almost sure convergence of the algorithms, a stronger result for the fixed $\mu$ algorithms.

If the ODE is locally stable, then the algorithm (1) is locally stable (indicating probable success of the adaptive scheme), while if (3) is unstable, then (1) is also unstable, and the adaptive algorithm fails. For instance, it is well known [4] that if the correlation matrix of the input process $E\{XX^T\}$ is positive definite, then (for small enough $\mu$) the parameter estimate errors of the LMS algorithm converge in distribution to a region about the origin. The same matrix $E\{XX^T\}$ appears



0-7803-0532-9/92 $3.00 

in our analysis as the linearization of $\hat{H}(W)$. Positive definiteness of this matrix implies local stability of the ODE, while a negative eigenvalue would imply local instability. We investigate, and give local stability conditions for the three signed variants of LMS. The condition for the sign-sign algorithm is new.

The relation between the adaptive algorithm (1) and the ODE (3) may be thought of as a type of "law of large numbers." To investigate how close the behavior of the algorithm is to the deterministic trajectory of the ODE, one desires a corresponding "central limit theorem." Consider the time scaled process $W_\mu(t)$. In section 2, the martingale central limit theorem is exploited to show that the error process

$$V_\mu(\cdot) = \frac{1}{\sqrt{\mu}}(W_\mu(\cdot) - W(\cdot)) \tag{4}$$

converges to a forced ODE that is driven by a sum of independent Brownian motions. Under some assumptions on the input and disturbance processes, the limit distribution is a Ornstein-Uhlenbeck process, with known mean and variance.

In practical terms, this convergence has two major implications. First, for a given algorithm, it is easy to calculate the parameters of the convergent distribution in terms of the properties of the inputs and disturbances, and hence to give a measure of the performance of the algorithm. Second, this allows a fair comparison between competing adaptive schemes, which we perform for LMS and its signed variants.

## 2. Principal Theorems

The problem may be formulated as follows:
Assumption (C.1) $\{Y_k\}$ is stationary , ergodic and there is a sequence of i.i.d. $E_3$-valued random variables $\{\psi_k\}$, independent of $\{Y_k\}$, and a measurable function $q : \Re^d \times E_1 \times E_3 \to E_2$ such that $U_{k+1} = q(W_k, Y_k, \psi_k)$, and $W_0$ is independent of $\{(Y_k, \psi_k)\}$. Define $P(U_{k+1} \in C|\mathcal{F}_k) = P(q(W_k, Y_k, \psi_k) \in C|\mathcal{F}_k) = \eta(W_k, Y_k, C)$ and assume that $H$ is integrable with respect to $\eta(w, y, \cdot)$ for each $(w, y) \in \Re^d \times E_1$. $\nu_Y \in \mathcal{P}(E_1)$ will denote the distribution of $Y_k$. Define

$$\bar{H}(w, y) = \int_{E_2} H(w, y, u)\eta(w, y, du). \tag{5}$$

Assumption (C.2) $\bar{H}$ is continuous in $(w, y)$ , and for $K \in \Re^+$

$$E\{\sup_{w:|w|\leq K}|H(w, Y_k, q(w, Y_k, \psi_k))|\} < \infty \tag{6}$$

$$E\{\sup_{w:|w|\leq K}|\bar{H}(w, Y_k)|\} < \infty. \tag{7}$$

$(E, r)$ denotes a metric space with associated Borel field $\mathcal{B}(E)$. Let $D_E[0, \infty)$ be the space of right continuous functions with left limits mapping from the interval $[0, \infty)$ into $E$. We assume that $D_E[0, \infty)$ is endowed with the Skorohod topology. Let $\{X_\alpha\}$ (where $\alpha$ ranges over some index set) be a family of stochastic processes with sample paths in $D_E[0, \infty)$ and let $\{P_\alpha\} \subset \mathcal{P}(D_E[0, \infty))$ be the family of associated probability distributions (i.e. $P_\alpha(B) = P\{X_\alpha \in B\}$ for all $B \in \mathcal{B}(E)$). We say that $\{X_\alpha\}$ is relatively compact if $\{P_\alpha\}$ is relatively compact in the space of probability measures $\mathcal{P}(D_E[0, \infty))$ endowed with the topology of weak convergence. The symbol $\Rightarrow$ will always denote weak convergence.

**Theorem 2.1** *Let $W_\mu(t) = W_{[t/\mu]}$, and for $K \in \Re^+$, define $\tau_\mu^K = inf\{t : |W_\mu(t)| \geq K\}$, and $W_\mu^{\tau_\mu^K}(\cdot) = W_\mu(\cdot \wedge \tau_\mu^K)$ define the "stopped" process. Assume C.1, C.2, and that $W_\mu(0) \to w_0$ in probability as $\mu \to 0$. Then for each $K$, $\{W_\mu^{\tau_\mu^K}, \mu > 0\}$ is relatively compact, and every limit point (as $\mu \to 0$) satisfies*

$$W(t) = w_0 + \int_0^t \hat{H}(W(s))ds \tag{8}$$

*for $t < \tau^K = inf\{t : |W(t)| \geq K\}$.*

Alternatively, we may assume: Assumption (C.2a) Define $\check{H}(w, y, z) = H(w, y, q(w, y, z))$. Let $Q = \{(w, y, z) : \check{H}$ is continuous at $(w, y, z)\}$. Assume that $\int\int I_Q(w, y, z)\nu_Y(dy)\nu_\psi(dz) = 1$, for every $w$ , and for $K \in \Re^+$

$$E\{\sup_{w:|w|\leq K}|H(w, Y_k, q(w, Y_k, \psi_k))|\} < \infty \tag{9}$$

$$E\{\sup_{w:|w|\leq K}|\bar{H}(w, Y_k)|\} < \infty. \tag{10}$$

**Corollary 2.1** *Assume C.1, C.2a, $W_\mu(0) \to w_0$ almost surely, and that the solution of (8) is unique. Then $\{W_\mu^{\tau_\mu^K}\}$ converges almost surely to $W^{\tau^K}$.*

All proofs may be found in [5]. Consider the scaled error process (4)

$$V_\mu(t) = \frac{1}{\sqrt{\mu}}(W_\mu(t) - W(t)), \tag{11}$$

where the scaling factor $\frac{1}{\sqrt{\mu}}$ expands $V_\mu$ to compensate for the time compression of $W_\mu(t)$. $G(w, y, u) = (H(w, y, u) - \bar{H}(w, y))(H(w, y, u) - \bar{H}(w, y))^T$ is the matrix that represents the deviation of $H$ from its smoothed version $\bar{H}$. If $H$ is square integrable with respect to $\eta(w, y, \cdot)$ for each pair $(w, y) \in \Re^d \times E_1$, we can define a smoothed version of $G$ as

$$\bar{G}(w, y) = \int_{E_2} G(w, y, u)\eta(w, y, du). \tag{12}$$

Averaging over all inputs yields

$$\hat{G}(w) = \int \bar{G}(w,y)\nu_Y(dy). \tag{13}$$

In addition to C.1 and C.2, we make the further assumptions:
Assumption (C.3) $H$ is square integrable with respect to $\eta(w,y,\cdot)$ for each pair $(w,y) \in \Re^d \times E_1$. $\bar{H}$ is differentiable as a function of $w$, $\bar{G}$ and $\partial_w \bar{H}$ are continuous, and for $K \in \Re^+$

$$E\{\sup_{w:|w|\le K}|H(w,Y_k,q(w,Y_k,\psi_k))|^2\} < \infty$$
$$E\{\sup_{w:|w|\le K}|\bar{G}(w,Y_k)|\} < \infty$$
$$E\{\sup_{w:|w|\le K}|\partial_w \bar{H}(w,Y_k)|\} < \infty$$

Define

$$\tilde{M}_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (H(W_k,Y_k,U_{k+1}) - \bar{H}(W_k,Y_k))\sqrt{\mu} \tag{14}$$

and

$$L_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (\bar{H}(W(k\mu),Y_k) - \hat{H}(W(k\mu)))\sqrt{\mu}. \tag{15}$$

There are a variety of different conditions (for example, mixing conditions on $\{Y_k\}$) that imply $\{L_\mu\}$ converges weakly to a (time inhomogeneous) Brownian motion. We simply assume this convergence. Thus the fourth assumption, (C.4), is that $L_\mu \Rightarrow L$.

**Theorem 2.2** *Assume C.1-C.4, that $W_\mu(0) \to w_0$ in probability, that the solution of (8) exists for all $t \ge 0$, and that $V_\mu(0) \to v_0$ in probability as $\mu \to 0$. Then $\tilde{M}_\mu \Rightarrow \tilde{M}$ where $\tilde{M}$ is a mean zero Brownian motion independent of $L$ with*

$$E\{\tilde{M}(t)\tilde{M}(t)^T\} = \int_0^t \hat{G}(W(s))ds$$

*and $V_\mu \Rightarrow V$ satisfying*

$$V(t) = v_0 + \tilde{M}(t) + L(t) + \int_0^t \partial_w \hat{H}(W(s))V(s)ds \tag{16}$$

## 3. Examples

In the various examples, we impose some common additional assumptions on the input and disturbance processes. These are not required by the theory. Rather, they are a way to find relatively simple expressions for the stability/instability of the ODE, and for the mean and variance of the corresponding Ornstein - Uhlenbeck process. These assumptions are:

- **E1** $\{U_k\}$ is a zero mean i.i.d. sequence with probability distribution $\eta(\cdot)$ and bounded density $f_u(\cdot)$ with $f_u(0) > 0$. The sequence $\{X_k\}$ is a stationary, ergodic sequence (with finite mean and covariance) of $\Re^d$ valued random variables independent of $\{U_k\}$.
- **E2** Assumption (E1) holds and that the components of $X_j = (X_{j1}, X_{j2}, \ldots, X_{jd})^T$ are i.i.d. symmetric, mean zero, variance $\sigma_x^2$ random variables for all $j \in \mathcal{Z}$.

Note that (E2) does not require independence of the vectors $X_j$. These, of course, are a very restrictive set of assumptions. However they will allow us to compare in a common setting the local stability/limiting distribution behavior of the four algorithms.

### 3.1 Sign-Sign Algorithm

The sign - sign algorithm, prized for its computational simplicity, has seen a resurgence of interest since its incorporation in a recent CCIT standard for adaptive differential pulse code modulation. Despite some efforts, a clear and simple test for stability of the algorithm has been elusive. The algorithm is

$$W_{k+1} = W_k - \mu \mathrm{sgn}(X_k)\mathrm{sgn}(X_k^T W_k + U_k) \tag{17}$$

where $W_k$ is the parameter estimate error, $X_k$ is a regressor of past inputs, $\mathrm{sgn}(X)$ applied to a vector is an element by element operation, and $U_k$ is a disturbance term.

Suppose that the $\{U_k\}$ sequence satisfies (E1). Define $y = (x, \mathrm{sgn}(x))$ or $Y_k = (X_k, \mathrm{sgn}(X_k))$. Then

$$-\bar{H}(w,y) = \mathrm{sgn}(x)(1 - 2\eta(-x^T w))$$

is continuous in $(w,y)$. Thus, conditions C.1 and C.2 (and hence Theorem 2.1) hold.

Let $F(\cdot)$ denote the distribution function of $X_1$. Then

$$-\hat{H}(w) = \int \mathrm{sgn}(x)[1 - 2\eta(-x^T w)]dF(x). \tag{18}$$

Since $f_u$ is bounded, we can show $\hat{H}$ is globally Lipshitz. Hence there exists a unique solution to the ODE which doesn't become unbounded in finite time. Therefore we do not need to work with the "stopped" processes. We linearize $\hat{H}$ about zero to obtain

$$-\frac{\partial}{\partial w}\hat{H}(0) = 2f_u(0)E\{\mathrm{sgn}(X_1)X_1^T\}.$$

For the "central limit theorem" results, note that

$$\bar{G}(w,y) = \mathrm{sgn}(x)\mathrm{sgn}(x^T)(1 - (1 - 2\eta(-w^T x))^2) \tag{19}$$

Assume (E2) holds, then $E\{\text{sgn}(X_1 X_1^T)\} = I$. Hence

$$\hat{G}(w) = I - E\{\text{sgn}(X_1 X_1^T)(1 - 2\eta(-X_1^T w))^2\} \quad (20)$$

or $\hat{G}(0) = (1 - (1 - 2\eta(0))^2)I = I$.

Recall that the Brownian driving term $L(t)$ is the limit of $L_\mu(t)$ of (15). At the equilibrium $w = 0$, $\bar{H}(0, Y_k) = -\text{sgn}(X_k)(1 - 2\eta(0)) = 0$. Similarly $\hat{H}(0) = 0$ for symmetric noise, which implies that $L_\mu(t) \Rightarrow L(t) = 0$

Hence, the limiting stochastic differential equation is:

$$V(t) = v_0 + \tilde{M}(t) - 2f_u(0)E\{\text{sgn}(X_1)X_1^T\}\int_0^t V(s)ds. \quad (21)$$

Under assumption (E2), the $V(\cdot)$ process "decouples" into $n$ independent components $V(t) = (V_1(t), V_2(t), \ldots, V_n(t))^T$ where

$$V_i(t) = v_{0i} + \tilde{m}(t) - 2f_u(0)E\{X_{1i}\text{sgn}(X_{1i})\}\int_0^t V_i(s)ds.$$

This is the general form of an Ornstein-Uhlenbeck random process. Define $\alpha = 2f_u(0)E\{X_1\text{sgn}(X_1)\}$ and $\sigma^2 = 1$. Then $V_i(t)$ is an asymptotically stationary Gaussian random process with mean zero, variance $\frac{\sigma^2}{2\alpha}$ and autocorrelation function $R_v(\tau) = E\{V_i(t+\tau)V_i(t)\} = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$.

Practically speaking, this means that for small $\mu$ we have the approximation $V_\mu(t) = \frac{1}{\sqrt{\mu}}(W_\mu(t) - W(t)) \approx V(t)$, where $V(t)$ has a $N(0, \frac{\sigma^2}{2\alpha})$ density, and $W(t) \approx 0$. Hence $W_\mu(t) = W_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu}{4f_u(0)E\{X_1\text{sgn}(X_1)\}})$ density.

## 3.2 Signed Error Algorithm

The signed error algorithm is similar to (17) but with the sgn function applied only to the error term

$$W_{k+1} = W_k - \mu X_k \text{sgn}(X_k^T W_k + U_k). \quad (22)$$

Emulating the above derivation (with $y = x$ or $Y_k = X_k$), one obtains $-\partial/\partial w\hat{H}(0) = 2f_u(0)\sigma_x^2 I$. Also $\alpha = 2f_u(0)\sigma_x^2$ and $\sigma^2 = \sigma_x^2$. Hence, as before we have, $R_v(\tau) = E\{V_i(t+\tau)V_i(t)\} = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$ and $W_\mu(t) = W_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu}{4f_u(0)})$ density.

## 3.3 Signed Regressor Algorithm

Applying the sgn function to only the regressor vector $X_k$ yields

$$W_{k+1} = W_k - \mu\text{sgn}(X_k)(X_k^T W_k + U_k). \quad (23)$$

With $y = (x, \text{sgn}(x)$ or $Y_k = (X_k, \text{sgn}(X_k))$, we may show $-\partial/\partial w\bar{H}(0) = E\{\text{sgn}(X_1)X_1^T\}$. Also $\alpha = E\{\text{sgn}(X_{1i})X_{1i}\}$ and $\sigma^2 = \sigma_u^2$. Then $R_v(\tau) = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$, and $W_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu\sigma_u^2}{2E\{X_1\text{sgn}(X_1)\}})$ density.

## 3.4 LMS Algorithm

Probably the most studied adaptive algorithm is the Least Mean Square algorithm [4]

$$W_{k+1} = W_k - \mu X_k(X_k^T W_k + U_k) \quad (24)$$

In this case $-\partial/\partial w\hat{H}(0) = -\sigma_x^2 I$. Also $\alpha = \sigma_x^2$, $\sigma^2 = \sigma_x^2\sigma_u^2$. Then $R_v(\tau) = \frac{\sigma^2}{2\alpha}\exp(-\alpha|\tau|)$ and $W_{[t/\mu]}$ has (approximately) a $N(0, \mu\frac{\sigma^2}{2\alpha}) = N(0, \frac{\mu\sigma_u^2}{2})$ for its stationary density.

## 4. Conclusion

The ODE results of [1] - [3] have been generalized to an almost sure result, even for the fixed stepsize algorithms. The theorems have then been applied to four popular adaptive filtering algorithms, leading to conditions for local stability (and in some cases, local instability) of the recursive algorithms.

## References


[1] A. Benveniste, M. Metivier, P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.

[2] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press Series in Signal Processing, Optimization, and Control, Cambridge, Massachusetts 1984.

[3] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. on Automatic Control* Vol. 22, No. 4, August 1977.

[4] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, Jr., "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proceedings of the IEEE*, Vol. 64, No. 8, pp. 1151-1162, August 1976.

[5] J. A. Bucklew, T. Kurtz, and W. A. Sethares, "Local Stability and Tracking of Adaptive Algorithms," submitted to *IEEE Trans. on Info. Theory*.